InterPARES Trust AI

Title	Annotated Bibliography - MA04
Working group code	MA04
Study title	The Role of Records and RM in Environments Where Trustworthy AI is the Focus
Status	Final
Version	
Writers	Matthew Hetu
Date	October 8, 2022

Annotated Bibliography

Agostinho, D., D'Ignazio, C., Ring, A., Thylstrup, N. B., & Veel, K. (2019). Uncertain Archives: Approaching the Unknowns, Errors, and Vulnerabilities of Big Data through Cultural Theories of the Archive. Surveillance & Society, 17(3/4), 422–441.

> This collaborative article dives into a discussion about information practices surrounding big data and data archives—largely focusing on the topics of biases, systemic errors, and ethical challenges in the field. The article offers an outline of the Uncertain Archives research collective and shows how cultural theories of the archives can be applied to the empirical filed of big data. This leads to a critique of archival reason, made possible by post-structuralist thought, feminist, queer, postcolonial, and critical race theories. Outlining challenges to the archives' capacity to produce truth, evidence and categorise human identity—leading to an uncertain definition of archives. The article then uses these theoretical approaches of cultural archives to take a critical approach to archival reason in the present day, calling on the works of Saidiya Hartman, Diana Taylor, Rebecca Schnieder, Ann Cvetkovich, Ann Laura Stoler, Sara Edenheim, Jack Halberstam, Michelle Caswell, Marlene Manoff, Marika Cifor, and Tonia Sutherland. In doing so, the study recognises the historical roots of current practices of big data (data hoarding, storing, leaking, and wasting)—big data is technologically new, but belongs in a long historical trajectory.

The article outlines three related conceptual lenses: *Unknown*/*Unknowable, error* and *Vulnerability.* Through these lenses the article thinks through the archives in terms of knowledge, power, and control, presenting the archives in terms of selection and interpretation allows for new epistemological and political implications when considered in terms of collection and use of big data. Using theories from performance studies the article paints an alternative perspective to conventional archival inscription and outlines the archives as something that overlooking the experience of women and queer people—similar argument are seen in transnational and postcolonial studies—what counts as a human subject in an archive? This shows that information collection is not a neutral pursuit, capture and exclusion of data has ethical consequences, big data in many ways extends this problematic nature of traditional archival reason. Datafication is embedded with means of uncertainty and risk. Big data for government and big companies is welcomed as a solution to deal with informational uncertainty—the promise of accurate calculations and prediction, yet they also frame big data as drivers of creativity and high-gain (Techno-capitalism approach). We see then that big data is not simply a rational apparatus, but a reflection of society's grappling with and fear of uncertainty. The article then outlines the unknown/mapping as a fundamental archival-technological function. Leading to a discussion of the black box coming to represent what we do not or cannot know. Presenting big data as creating new forms of knowing, but also new forms of unknowing.

The article then moves into a discussion of error, using a psychoanalytical approach and placing error in big data in a long tradition of error studies from psychology and economics to engineering. It outlines some of the ethical concerns regarding error in big data: big data as political sites of information not objective statements of truth, error as a crucial computational function, overcoming error, error as challenging regimes of control, and more. Finally, the article moves into discussing vulnerability. Outlining vulnerability as a shared condition of the human and non-human but says that not all subjects are equally vulnerable. Though this approach pulls on a long history of cultural studies it is still an anthropocentric approach. The article asks "how are people and communities affected differently by big data archives?" Yet, this approach seems to dismiss or discredit the nonhuman and ecological—something that seems problematic in an era of climate emergency. The article discusses BIPOC and the continuation of the commodification and violence on black bodies that makes these people more vulnerable to current big data archives, thus recentering the discussion of archival tension between capture and exclusion—big data is embedded in historical relations of racial capitalism. Going forward, the article claims, we must revisit the archives and archival theory and engage with voices from other forms of knowledge production to understand and live in a time of, and become critical of, big data and an era of data surveillance.

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias. ProPublica, May, 23, 2016.

This article investigates the algorithms used in the USA to predict future criminal acts and biases the algorithms have against people of colour. ProPublica starts by outlining what these scores (risk assessments) are and how they are becoming increasingly common in courtrooms across the USA. These scores are used to inform decisions throughout the justice system. The article expresses that "In Arizona, Colorado, Delaware, Kentucky, Louisiana, Oklahoma, Virginia, Washington and Wisconsin, the results of such assessments are given to judges during criminal sentencing."

In response to the increasing use of these risk assessment scores an investigation led by ProPublica as part of a larger examination of the effects of algorithms in American life was undertaken. They obtained the risk scores of more than 7000 people from Broward Country, Florida in 2013 and 2014 and checked to see how many of these people were charged with new crimes over the next two years. The results showed that scores were remarkably unreliable (only 20% of those predicted to commit violent crimes actually did. These results also showed that there were significant racial disparities in the scores. The formula was particularly likely to falsely flag black defendants as future criminals, wrongly labeling them this way at almost twice the rate as white defendants. Of note, there is also a separate article outlining their statistical process to the investigation: "How We Analyzed the COMPAS Recidivism Algorithm".

The article then touches on a larger issue within algorithms—that of the 'black box". The calculations used in the score are not publicly disclosed, leading to a system where the defendants rarely have the opportunity to challenge their assessments. The article then points out that judges have mentioned using these scores in their sentencing decisions. Leading to a call for more transparency and more investigation into biases that are present in these algorithms, and more regulations on how they can, and should be used in the court.

Barfield, W., Pagallo, U., & Elgaronline. (2018). Research handbook on the law of artificial intelligence. https://doi.org/10.4337/9781786439055

This book consists of 25 chapters written by leading legal and AI experts throughout the USA, Asia, And the European union. This book is wide in scope—touching on topics in private, corporate, criminal, and constitutional law. It focuses on concepts of regulation, rights, intellectual property, and applications of AI within juridical contexts. The text also makes clear that notions of AI law and regulations need to be international in scale to address the scope and transnational impact of many algorithm and AI systems. It stems out of a need to address how emerging AI technologies and systems are challenging and complicating many areas of law and legislation, yet there are no major regulatory schemas for AI use, even though AI is involved in almost all aspects of society. The scholars also make clear the difficulty in legally defining what counts as AI, protecting the rights of the parties involved in AI use and creation and the concepts of AI legal personhood status. The text then seeks to be a step forward in the current legal framework for thinking about law and AI, and lead future discussion and directions by leading scholars in the field.

Of particular note is Chapter 19 "Artificial Intelligence and the Creative Industry: New Challenges for the EU Paradigm for Art and Technology by Autonomous Creation" by Madeleine de Cock Burning. This chapter discusses how robotics and other computer programs are challenging EU's regulatory framework and the policy domain. This particular article focuses on the field of intellectual property protection for autonomous creation—and the challenges involved when a machine is the creative agent of a work. In particular what aspects of protection would these work and outputs have under EU's current copyright framework. By outlining advancements in AI and a reviewing copyright law and policies the chapter clearly shows that policies need to be re-evaluated to address challenges that emerging AI brings to the legal framework (both within copyright law and beyond)—outlining the desire for and challenge of creating a future proof legal framework.

Bunn, J. (2020). Working in Contexts for Which Transparency Is Important: A Recordkeeping View of Explainable Artificial Intelligence (XAI). Records Management Journal 30(20): 143–53. https://doi.org/10.1108/RMJ-08-2019-0038.

This article introduces the topic of explainable artificial intelligence (XAI) and outlines the outcome of an interdisciplinary workshop on the topic—reflecting on XAI through the frame of the recordkeeping profession. Bunn takes a reflective approach to the topic and also outlines a historical trajectory for XAI. Bunn's introduces XAI as "shedding light on opaque machine learning in contexts for which transparency is important." The article stems from the shared spaces of transparency between AI and recordkeeping and tries to find a common ground between the two field—linking the two with notions of transparency, accountability, fairness, social justice, and trustworthiness. Bunn's then outlines how notions of opaque AI technology is causing the need for a rethinking of recordkeeping practices and what we can capture to make records around these technologies more transparent, arguing that these AI technologies are becoming agents of transactions and asking, "what does evidence(records) of these transactions look like?".

The article seeks to then create a space for interdisciplinary conversation and outlines the results of a workshop held between the National Archives and the

Human Computer Interaction. In this conference emerged the idea of HeXAI or human-centeredXAI. In turn, questioning the metaphor of the black box and moving to a methodology with more transparency and agency for the user and recordkeeper. In doing so, the article discusses a shift in the conversation to the human need for explanation, asking questions like: When do we need to offer an explanation? How detailed does it need to be? And why is explanation needed? In raising these questions Bunn's turns to notions of fairness, accountability, and transparency. The article attempts to make a connection between these notions in AI and these notions in recordkeeping pulling from EU Commission High Level Expert Group on AI and also InterPARES Trust as examples of these conversation. Finally, the article concludes by raising questions in terms of the records created around these systems asking us to think about what record are created around creating AI systems, what records are created of the decisions and impacts of the systems, do they meet legal provisions, and do they meet the required standard of quality.

Chabin, M.-A. (2020). The potential for collaboration between AI and archival science in processing data from the French great national debate. Records Management Journal, 30(2), 241–252. https://doi.org/10.1108/RMJ-08-2019-0042

This article was set in the context of the French great national debate in 2018, where proposals were gathered in the form of public meetings and a centralized digital platform—producing a considerable amount of data. The government, due to the urgency of the situation chose to use AI and private companies to process all the data in a timeframe of two to three weeks. In turn, the article seeks to offer a critical perspective of the types of algorithms that were used in the great national debate and how incorporating archival expertise may have enriched the information presented from AI.

The data was collected in two distinct ways 1)delocalized operations and local public meetings 2) national platform consisting of a digital questionnaire with four themes and 84 questions. This consisted of nearly 20,000 citizen notebooks, more than 27,000 letters and emails address to the great debate mission, 10,000 minutes of meetings, 1.9 million contributions deposited on the digital platform. The data was processed in two sections, OpinionWay delt with the closed-ended questions and outsourced the analysis of the open-ended questions to Qwam Content Intelligence. Roland Berger Company and Cognito and Bluenove delt with the processing of the other forms of contributions. The paper materials were pre-processed by the Bibliotheque Nationale de France to make them into materials that could be used by algorithms. The choice to use AI was explained by the volume of data and the time frame of the political calendar. For the platform data the closed-ended questions are processed by counting. However, the open-ended questions were analyzed by "coherent word groupings" by QWAM text analytics. For the other data the key concern was the development of a lexicographic reference framework. Both of these approaches were subject to criticism from others in the field of AI. Largely calling for more transparency of the processing operations carried out. The data was also processed by supervised algorithms, while unsupervised learning technologies are more effective. Sematic vs cognitive approach to algorithms. Semantics makes the machine learn keywords—the algorithm then searches, groups and counts based on these terms. But by presupposing the themes we may miss out on interesting findings. The cognitive approach seeks to extract themes or categories without any pre-set reference. Therefore, the categorization does not require "supervision".

There was no archival expertise included in the project of the great debate in terms of processing and creating the data. Two major issues stem from this: the first is the dismissal of other descriptive elements besides the text (origin, date, structure, etc.) and the second is that some subjects were left out of the analysis. The consultation gave the citizens the freedom to participate or not therefore it is not representative as it encouraged the participation of people with higher formal education. The themes and categories also do not allow any room for interpretation or context. Arguments then arise pushing for a diplomatic approach by archival experts. To determine if the processing of the data could be improved by taking into account the formal elements. For example, the processing did not account for dates and the time period of the response could greatly provide greater information. The titles of the responses were open too and are deserving of study. The data is also a public record, yet there has been no mention of archiving. Suggesting that records managers and archivists should take a more active role in the creation of such records.

Cheatham, B., Javanmardian, K., & Samandari, H. (2019). Confronting the risks of artificial intelligence. McKinsey Quarterly, 1-9.

This article offers a surface level introduction to the risks and risk mitigations strategies that businesses face when deploying AI tools and systems. It outlines potential negatives of using such systems such as, privacy violations, discriminations, accidents, and manipulation of political systems. These outcomes could have disastrous repercussions, but the article asks the question "How do leaders in business mitigate these risks?" The article is very much grounded in the context of using AI in business and an economic frame. It outlines five potential risks. The first, data difficulties such as inadvertently using or revealing sensitive information hidden among anonymized data. This is of particular importance due to emerging privacy legislation such as GDPR and CCPA and changes to BC FIPPA. The technology trouble, technology and processing issues can negatively impact the performance of AI systems. The third risk is security snags, the potential for fraudsters to exploit the data collected by AI systems. The fourth risk is models misbehaving and doing things like delivering biased results. Finally, the article discusses interaction issue in the interface between people and the machines—leading to human errors. The article functions as a thought piece pushing for more action in terms of policing, engagement and ethical use of AI systems in the work place and argues for more research and though to be done on the subject.

Citron, D. K. (2007). Technological due process. Wash. UL Rev., 85, 1249.

The article seeks to engage in crucial conversation about protecting our due process values in a world of automation. Automation has enormous potential, but Citron pushed that we must be careful and aware of the errors it can produce and consider protecting individual's interests in a fair and transparent ways. The article outlines the growing influence and size of executive administrative agencies. This administrative state faced serious criticism such as agency capture and the ossification of rulemaking. The twenty-first century automated decision-making systems bring radical change to the administrative state. There are many benefits of introducing automation into such a system such as cost savings and reducing physical hassle. Yet, it is also a risk for dismantling critical safeguards and laws. The article then offers a framework for administrative and constitution law designed to address the challenges of the automated state. It does so in three parts, the first describes how automated systems are built and the varying ways that policies are embedded and (often) distorted in their code. Part two discusses how automation jeopardizes procedural protections. Finally, part three articulates a new model of technological due process offering a systematic way for an agency to approach using automation. Drawing on rules-versus-standards literature to provide a systematic approach to deciding between automation and human discretion.

Cohasset Associates, & ARMA International. (n.d.). 2019 Information Governance Benchmarking Report. Retrieved from https://armai.informz.net/ARMAI/pages/Cohasset_Benchmarking_Survey_2019

This is a benchmarking report published by Cohasset associates and ARMA International. It focuses on the practices of records and information management (RIM) and information governance (IG) and the shift between the two areas. In turn, the report documents the critical evolution to IG by business dynamics, legal implication, and technology innovation—providing data on information lifecycle management practices and process with a focus on electronically-stored information. The survey reports on three key categories: the state of IG advancement, achievements and the obstacles resulting from and impacting IG, and actions strategies that facilitate effective information lifecycle management. The survey was conducted using a web-based survey tool. Nearly 900 survey responses were recorded during February and March 2019. Over 14,000 responses since its inception and is considered the state of IG. Of particular note to our research is the discussion of automated processes and tools that enable organizations to manage information over its lifecycle. This comes up in the technical challenges, deletion, and technology advancement sections of the survey.

d'Alessandro, B., O'Neil, C., & LaGatta, T. (2017). Conscientious Classification: A Data Scientist's Guide to Discrimination-Aware Classification. Big Data, 5(2), 120–134. https://doi.org/10.1089/big.2016.0048

Automation is increasingly moving from the hand to the brain (intellectual vs manual tasks). The article mentions many benefits to using machine learning, but also cautions that subtle and ugly truths have also come to the forefront. One such major issues is reinforcing historical systemic biases. The article asks how can the ethical data scientist do better? The researchers approach this question in two sections the first discusses the root causes of discrimination. The second offers a broad survey of discrimination measures and discrimination-aware data mining methods. The article is presented with the practicing data scientist in mind with the intent to spur data scientists to action.

D'Alessandro starts with a discussion of exactly what discrimination is—offering conceptual and legal definition of discrimination and the means to actually measure it within data or within a machine learning system. The article focuses primarily on classification and ranking systems. It then moves into a discussion of causality and disparate treatment. This section outlines the paradox that including protected attributes in classification runs tremendous risk of liability under disparate treatment doctrine and legal risk, but lack of knowledge of these attributes reduces one's ability to detect and avoid disparate treatment. It then discusses statistical regression model for measuring disparate impact and other models that can be used to measure discrimination measures. The article outlines two sources of discrimination: 1) data issues 2) Misspecification. Data issues are the most straightforward: "discrimination in, discrimination out". This could include things like sample bias. Misspecification is a common concept in statistics, which can be described as the functional form or feature set of a model under study not being reflective of the true model.

The article then concludes its discussion of discrimination by debating the data scientist's role in creation. The greatest error one can make during evaluation is to not test an algorithm for potential discriminatory behavior. Also, important to have appropriate feedback loops and making good judgments on when to involve human experts in the decision-making process. Keeping humans in the link improves objectivity and nuanced flexibility The article then proposes a discrimination-aware auditing process. Using this system, the article proposes several "discrimination aware" unit tests to help guide future development. Then offering a number of case studies outlining some of the problems and techniques they have previously discussed. The article does not propose any new methodological techniques, it hopes that this survey might serve as a useful first guide in both discrimination measurement and discrimination-aware data mining for those active in the field.

Frendo, R. (2007). Disembodied information: Metadata, file plans, and the intellectual organisation of records. Records Management Journal, 17(3), 157–168. https://doi.org/10.1108/09565690710833062

Frendo provides a critical review of the literature discussing discrete electronic metadata capture and the debate between automation and traditional classification. The article presents a recognition that contextual structures and relationships cannot at present be automated, natural language processing capabilities are poor, and metadata can easily become decoupled from "disembodied" discrete units of information. Discrete metadata capture has been developed in the context of commercial transactions rather than information management.

The article pulls from works such as Bruno Delmas' "Archival science facing the information society" to discuss the issue of when and how a record's metadata should be created and captured. With the majority of scholars such as Bearman, Margaret Hedstrom and David Wallace agreeing that "descriptive practices originating in a computer systems environment, as well as the descriptive methods used by data archives, fall short of what is needed because they focus on data structures and content with insufficient regard for the contextual information needed to define and understand electronic records." There is a fundamental incongruence between the intellectual structure offered on the one hand by file plans and directory structures, and on the other by metadata which is generated individually for discrete units, and which provides no form of association between records other than identity of an attribute. An additional shortcoming the article mentions, is the difficulty of maintaining persistent links between metadata and record content—pulling from researchers attempting to develop metadata creation models for the preservation of digital records. The article then moves to discuss the possible implications and implementations of metadata structures in recordkeeping systems. Arguing that the adaptation of any system seems to be determined by established practice or cost-effectiveness. Frendo pushes that evidential, integrity, authenticity, and robustness are (or should be) of equal importance. Claiming then, that replacing traditional classification structures, with metadata specific to individual transactions may seem like a convenient use of automation. However, discarding human-centric approaches risks "relinquishing those attributes of records which lend them their significance, both present and future."

Marcus, G. (2018). Deep Learning: A Critical Appraisal. ArXiv:1801.00631 [Cs, Stat]. Retrieved from http://arxiv.org/abs/1801.00631

The article presents ten concerns for deep learning, and suggests that deep learning must be supplemented by other techniques if we are to reach artificial general intelligence. Before getting into these challenges the article asks, "is deep learning approaching a wall?" This leads Marcus to offer a crucial reflection on the field made both for researchers and AI consumers. The article then offers a brief outline of deep learning—essentially a statistical technique for classifying patterns, based on sample data, using neural networks with multiple layers. It goes on to explore the limits on the scope of deep learning: such as problems of contrapositives and generalizations and offers ten challenges in the current deep learning systems.

The first challenge is that deep learning is data hungry. It works best when there are thousands, millions or even billions of training examples. In problems where data is limited, deep learning is often not the ideal. There is also limited capacity for transfer and the patterns extracted by deep learning are often more superficial than they initially appear. The third challenge is that there is no natural way to deal hierarchical structures. This section pulls from linguist theory—notably the works of Noam Chomsky. When a complex hierarchical structure is needed a core problem occurs due to the fact that deep learning learns correlations in a "flat" manner—every feature is considered on equal footing. The article then goes on to discuss numerous other challenges the field is facing such as: struggle with open-ended inference, the field is not sufficiently transparent (this section is of particular importance when considering notions of explainable AI and leads to potential liability issues when using deep learning and can also lead to serious issues of bias), not well integrated with prior knowledge, cannot inherently distinguish causation from correlation, presumes a largely stable world, "spoofability" of deep learning systems, and being difficult to engineer with. The article then enters a discussion of these issues and states that the real problem lies in the misunderstanding what deep learning is, and is not, good for, mentioning that there are also potential risks stemming from the "hype" surrounding deep learning. The article then pushes for a reconceptualization of the field as simply one tool among many and offers other tools/areas to consider in tandem, such as unsupervised learning, symbol-manipulation, insight from cognitive and developmental psychology.

Miracchi, L. (2019). A competence framework for artificial intelligence research. Philosophical Psychology, 32(5), 588–633. https://doi.org/10.1080/09515089.2019.1607692

This article moves away from the common focus of building AI tools and applications and instead focuses on building a genuinely intelligent artificial agent. It offers a theoretical and methodological framework to provide new avenues for research. Miracchi focuses in on the term *Artificial Minded Intelligences* (AMIs) and outlines three main criteria that should be considered: 1)it should show how we can directly empirically ask the key question. Also, productively break down the key question into sub-questions, which may be approached to some extent independently by a collaborative team of researchers with different areas of expertise. 2) it should not commit to a particular technical approach: it should be general enough to encompass explicit symbolic, dynamical, and neural-net approaches – including deep learning – and potentially suggest useful ways of integrating and developing these approaches.

The article asks the question, how might artificial processes give rise to minded intelligence, both in the general case and in cases of specific mental kinds, such as perception, knowledge, language comprehension, and goal-directed action. It starts by arguing that intelligence and related mental properties should be treated as distinctive higher-level properties of artificial systems. The second section focuses on developing a new way of understanding in virtue of explanations in cognitive science, especially explanations of how mental kinds like consciousness, perception, and intention obtain in virtue of neural, computational, bodily, and environmental processes. It is meant to be a relatively neutral framing of the goal of AMI research, the study of how to build an artificial system with minded intelligence. In discussing these topics, Miracchi outlines concepts of definitive methodology, generative methodology, competence framework, robustness, flexibility, and autonomy. They then offer a general approach to taking mental intelligence seriously as a direct object of AI investigation—opening up the space for future research about mental intelligence and related topics.

Nikzad–Khasmakhi, N., Balafar, M. A., & Feizi–Derakhshi, M. R. (2019). The state-of-the-art in expert recommendation systems. Engineering Applications of Artificial Intelligence 82, 126-147.

This article is spurred from an increase in the amount of digital information and multimedia content, leading to more difficult searches and demands from user. It Outlines what Information Retrieval (I R) is as a form and framework for finding and accessing information. This is applied in many applications and social networks. The work outlines three approaches for recommendation systems: collaborative recommendation systems, content-based filtering, and hybrid recommendation. It then goes on to summarize the history and influence of recommendation systems and outlines what an expert recommendation system specifically is saying, "an expert recommendation system takes the users' query firstly, next it gathers the past reputation of experts, then it classifies expertise into a subject classification schema, and finally provides a ranked list of experts that their expertise matches most closely to the user's query."

The team offers two key definitions for what preciously it means when it says expert: "Definition 1. User ui is called an expert if and only if his/her score is higher than the threshold θ, as described in Eq. 2. expert recommendation problem is initiated as finding a ranked list of experts $y ' i$ from list of features xi based on training dataset." The article outlines both content-based Information

Retrieval (CBIR) and Social Graph-based Information Retrieval (SGBIR). It also outlines both general purpose applications and specific purpose applications of expert recommendation systems provide services that are limited to particular topics or domains. One of these domains is health-care area. The team then goes on to investigate and classify the state-of-the-art in expert recommendation systems and outlines a series of evaluation metrics: Precision, F-measure, Mean Average Precision, Root Mean Square Error, Discounted Cumulative Gain, and then a comparison of systems in undertaken using these parameters. In doings so, the article proposed a procedure for an expert recommendation system, provides an overview of current systems and the advantages and disadvantages, and expressed evaluation metrics and existing challenges to the field and future research.

Osoba, O., Welser, W., IV, & RAND Reports. (2017). An intelligence in our image: The risks of bias and errors in artificial intelligence.

This article seeks to explore the consequences and risks of our increasing dependence on AI. The report outlines some of the shortcoming of algorithmic decision making and identifies problems surrounding algorithmic error and bias. It was written for decisionmakers and implementers in mind and was constructed by RAND Ventures—a research organization that develops solutions to public policy challenges to help make communities throughout. The goal outlined in the report is to explain the risk associated with uncritical reliance on algorithms.

The report starts by offering a definition and evaluation of algorithms, offering a discussion of the often opaque, uniformed understanding of algorithms in public discourse. Osoba argues that the opacity of algorithms makes it harder to judge correctness, evaluate risk, and assess fairness in social applications. Next, a brief history and discussion of examples of bias and misbehaving algorithms is offered, with particular empathise given to a case study on artificial agents in the criminal justice system.

The final section outlines factors and remedies of these machine biases. Such factors discussed include: the paradox of artificial agency, sample-size disparity, hacked reward functions, cultural differences, confounding covariates. It then moves to discuss potential remedies: causal reasoning algorithms, algorithmic literacy and transparency, and personnel approaches. The report outlines that response to unregulated artificial agents tends to be of three broad types: avoiding algorithms altogether, making the underlying algorithms transparent, or auditing the output of algorithms. Finally, of note the concluding discussions mentions the anthropomorphism that is often used surrounding the topic of AI. I found this particularly refreshing and useful as many conversations seem to overlook this anthropomorphism and I think it is deserving of thought and criticism—especially, as this report mentions, surrounding the conversation of algorithmic bias.

Rogers, C. (2019). From time theft to time stamps: Mapping the development of digital forensics from law enforcement to archival authority. International Journal of Digital Humanities, 1(1), 13–28. https://doi.org/10.1007/s42803-019-00002-y

Rogers offers a comparison of digital forensics and archival science and digital preservation. Presenting a brief investigation in the overlapping and shared histories and legacies of the two disciplines. Stating that both fields are concerned with discovering, understanding, describing, and presenting or making accessible digital materials. The purpose of digital forensics is predominantly in services of legal evidence and literature surrounding the field is often highly technical (computer science and mathematics). However, Rogers mentions that there have been recent calls for digital forensics to be situated within a social and theoretical framework. Rogers argues that the conceptual underpinning of the field can be examined through the lens of archival science, diplomatics, and the law.

The paper traces the chronological development of digital forensics from its evolution in the 1980s to the present day. Focuses on issues that shaped the field, such as society's increasing reliance on computer technology, collaborative approach by legal personnel, law enforcement, and IT specialists, and the spread of digital forensic from the law enforcement to other domains. By presenting this history, Rogers outlines the shared legal context of digital forensics and archival science. Pulling from diplomatics (Duranti) and digital forensics literature (Palmer) and scholars of the history of digital forensics (Charters, Pollitt, and Garfinkel). The article then points of parallels between the fields and shows the potential for an interdisciplinary comparison of digital forensics and archival diplomatics and their shared values. Arguing for shared benefits from incorporating concepts from digital forensics (authentication, reproducibility, non-interference, and minimization, laws of association, context, access, intent, and validation) into archival practice.

Stanford University. (2016). Artificial Intelligence and Life in 2030: One Hundred Year Study on Artificial Intelligence; Report of the 2015 Study Panel. Retrieved from https://ai100.standord.edu/sites/default/files/ai_100_report_0916fnl_single.pdf

This study was launched in the fall of 2014 and is a long-term investigation of the field of Artificial Intelligence and its influences on people, their communities, and society. A panel is formed every five years to assess the current state of AI. This particular study focuses on a typical North American city and is meant to highlight specific changes affecting the everyday life and mundane. The focus is then on eight domains: transportation, healthcare, education, low-resource communities, public safety and security, employment and workplace, home/service robots, and entertainment. In each domain, even as AI continues to deliver important benefits, it also raises important ethical and social issues, including privacy concerns. The article also outlines and discusses emerging AI research trends such as, large-scale machine learning, deep learning, reinforcement learning, robotics, computer vision, natural language processing, collaborative systems, crowdsourcing, game theory, internet of things, and neuromorphic computing.

Of particular value to our study is the section on AI policy. Public policies should help ease society's adaptation to AI applications, extend their benefits, and mitigate their inevitable errors and failures. In terms of policy the study panel offers three general policy recommendations:1) defining a path toward accruing technical expertise in AI at all levels of government 2) remove the perceived and actual impediments to research on the fairness, security, privacy, and social impacts of AI systems, and finally, 3) increase public and private funding for interdisciplinary studies of the societal impacts of AI. The article also outlines a number of legal considerations in regards to AI, saying it has the potential to challenge any number of legal assumptions in the short, medium, and long term. Legal consideration is given in terms of liability (criminal), agency, and certification, labor, and taxation. Showing that legal deliberation plays an important role in developing, enforcing, and implementing AI advancements.

Tan, Y. C., & Celis, L. E. (2019). Assessing social and intersectional biases in contextualized word representations. In Advances in Neural Information Processing Systems (pp. 13230- 13241).

This article seeks to investigate word embeddings such as word2vec and GloVe and their exhibiting of social biases, including gender bias and racial bias. Claiming that these biases are extremely concerning, as word embeddings form the foundation of most language systems. The work hopes to extend and expand on analysis of contextual representations with respect to social and intersectional biases. It departs from previous work by including the nominative (she), accusative (her), prenominal possessive (her) and predicative possessive (hers) inflections of personal pronouns, and also include the non-gendered or collective pronoun they—though their analysis still seems to function in a male-female binary.

They approached each sentence and incremented a count for female pronoun occurrence, if there are any in the sentence, and then a count for the pro-stereotypical and anti-stereotypical associations with occupation words (the same counts are then done for male and non-gendered pronouns). The text then pull from previous works (Caliskan and May) and use embedding association tests. However, Tan extends these tests to contextual word representations and introduces new embedding association tests to target race, gender, and intersectional identities. In doing so, the article shows that standard contextual have significant gender bias, extends existing tests to contextual word models and indicates social bias and racial bias in said models. In these tests, Tan shows the need for using both sentence encoding and contextual word representation. The article then points to future direction for studies in the field in particular investigating how and why the encoding of bias may differ across both model size and model layer.

Thibodeau, K. (2018). Computational Archival Practice: Towards A Theory for Archival Engineering. 2018 IEEE International Conference on Big Data (Big Data), 2753– 2760. https://doi.org/10.1109/BigData.2018.8622174

This article introduces the concept of archival engineering. Thibodeau states that archival engineering can be differentiated from archival science using Henry Petroski's simple assertion, "Science is about knowing, engineering is about doing." Archival engineering is then a systematic application of archival science to deliver optimal value.
It askes two fundamental questions: 1) What is done with archival resources? 2)What are the benefits of archival engineering for increasing and improving knowledge of the past? In asking these questions, the article first explores what is involved in knowing "the past". In turn, exploring how archival science can be applied to contribute to the production and improvement of this knowledge of the past. Within this critical discussion of the concept of "the past" Thibodeau introduces the concept of target pasts and constructing said target pasts. Entering into a discussion and definitions of purview, historical context, total context, materials and tokens.

The article outlines the place that archival engineering could take in this saying that it "offers the potential for improving the construction of target pasts and the evaluation of the results by building on concepts in archival science, expanding them to a broader scope, adapting them to encompass unprecedented aspects of digital information, facilitating automated processing, and enabling verification through quantitative testing." The article then outlines key terms and definitions such as *record* and archival concepts, and instead introduces the idea of an archival token. An archival token is a type of conceptual object. Thibodeau offers the following definition, "an archival token is a token that represents one or more objects from a former time in an advantageous manner because of its proximity to its referent both temporally and contextually." The article then goes on to discuss evaluation criteria such as notions of objectivity, translucency, and richness for archival engineering application and improving the construction of target pasts. Concluding that there is potential for archival engineering to increase and encourage the value of archival science.

Trace, C. B., & Francisco‐Revilla, L. (2015). The value and complexity of collection arrangement for evidentiary work. Journal of the Association for Information Science and Technology, 66(9), 1857–1882. https://doi.org/10.1002/asi.23295

Stems from the contrast between searching for evidence of a particular event, story, or scenario through general-purpose databases and search engines vs. the primary aim of archivists, and archival systems to enhance the long-term value of existing materials as evidence and support their interpretation by users. The goal of the Augmented Processing Table (APT) project is to enable archivists to manage increasing volumes of data and, in the process, to continue to facilitate, and indeed augment, those aspects of the curation workflow that support evidentiary work. The article provides an in-depth discussion of a study of archival curation practices involving both paper and digitized images. Touching on issues (clearing backlogs of unprocessed archival collections) and also proposes new methodologies and outcomes. Of particular note is how human-computer interaction can be used in the field of archives.

Upward, F., Reed, B., Oliver, G., & Evans, J. (2013). Recordkeeping informatics: Re‑figuring a discipline in crisis with a single minded approach. Records Management Journal, 23(1), 37–50. https://doi.org/10.1108/09565691311325013

The article's goal is to highlight the widespread crisis facing the archives and records management profession and to propose recordkeeping informatics, a single-minded disciplinary approach, as a way forward. The paper follows an Australasian perspective on the nature of the crisis besetting archives and records management professions as people struggle to adjust to digitally converged information systems. It presents that recordkeeping informatics as an approach for refiguring thinking, systems, processes and practices as people confront ever increasing information. The project started in 2008 when the group discussed the need for a new text to support records and archives information. The article also makes a distinction of recordkeeping as a one word description of the processes by which we create, capture, organise and pluralise records; and record keeping (two words) as a way of referring to the keeping of records as physical things.

Recordkeeping informatics covers the way we capture, archive and disseminate recorded information as evidence using modern communication and information technologies. It provides a bridge between records managers, archivists and information systems. The article also outlines records continuum and metadata as basic building blocks stating that a single-minded approach focuses on the way recordkeeping informatics can use a records continuum approach and recordkeeping metadata as ways of striving to help others bring order to the chaos of recordkeeping. The article then describes recordkeeping informatics relationship with other systems such as metadata schemas, organisational culture, and business process analysis, and access consideration. The article presents a view of recordkeeping informatics as a disciplinary base for human-based action and control. Encouraging critical thinking about recordkeeping informatics and the role that it can play in managing records and our interactions with information.

Upward, F. (2019). The monistic diversity of continuum informatics: A method for analysing the relationships between recordkeeping informatics, ethics and information governance. Records Management Journal, 29(1/2), 258–271. https://doi.org/10.1108/RMJ-09-2018- 0028

This paper aims to support an advance towards networked cohesion based on informatics. Upward's states that new regulatory approaches will have to manage *monistic diversity*, and pulls from continuum thinking and approaches the topic through studying things in motion as part of evolutionary processes. In doing so, the article seeks to connect thought, action, ethical information governance, and situates many operations under the joint term informatics. This line of thinking is situated within notions of the expansion in the continuum of records and recorded information and increasing notions of digitization and asks what can be done to mediate the disruptions in regards to authoritative information resource management.

Upward's argues that the continuum of recorded information is becoming more difficult to govern, and seeks to open up ideas about how to manage monistic diversity by looking at what it can mean to say "all is archive" as a base for developing continuum informatics as an integration tool. In introducing the idea that all is archive, the article presents archive as something that is innovative, complex and full of connections. However, Upward's warns that unless mediative factors are introduced new business models will continue to extend chaos in a world of digital content—hence the need for an ethical compass of sorts. Informatics offers a way of approaching the modern need for the convergence of disciplines without interfering with the continued development of specialisations—it does not interfere with expanding and emerging diversity. The article then thinks though internal and external mechanics of archival formation processes, concluding with the notions of cyber-maturity. Continuum informatics, according to Upward, will focus on the disruptions to the field and strive to be just as innovative but with an appreciation of the importance of mutual reciprocity and association.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... & Polosukhin, I. (2017). Attention is all you need. In Advances in neural information processing systems (pp. 5998-6008).

This article offers a technical review of an alternative model to using recurrent neural networks, long short-term memory and gated recurrent neural networks in particular. These recurrent networks have been firmly established as state-of-the-art approaches in sequence modeling and transduction problems. These recurrent models are typically used to factor computations along the symbol positions of the input and output sequences. They then generate a sequence of hidden states as a function of the previous hidden states.

Attention mechanisms have become an integral part of compelling sequence modeling and transduction models in various tasks. However, in all but a few cases, such attention mechanisms are used in conjunction with a recurrent network. This article proposes a model (the Transformer) that relies entirely on an attention mechanism to draw global dependencies between input and output. Allowing for more parallelization. This model follows the overall architecture using stacked self-attention and point-wise, fully connected layers for both the encoder and decoder. Mult-head attention is used in three primary ways: in the "encoder-decoder attention" layers, the encoder contains self-attention layers, similarly, self-attention layers in the decoder allow each position in the decoder to attend to all positions in the decoder. Self-attention is used because of three considerations: total computational complexity per layer, amount of computation that can be parallelized, and the path length between long-range dependencies in the network. Thus, presenting a model that can be trained in less time than a traditional recurrent model and can hopefully be applied to other sequential problems too.

Winters, J., & Prescott, A. (2019). Negotiating the born-digital: A problem of search. Archives and Manuscripts, 47(3), 391–403. https://doi.org/10.1080/01576895.2019.1640753

This article explores the limitation of search focused born-digital archives and seeks out possible approaches to an alternative such as the linking of files. Pulling from contemporary example from journalists, data leaks, and emails—exploring the challenges of our over-reliance on keyword searching. The article starts by a historical review of tools and methods to assimilate masses of new data and outlines Google's role in making key word searching the norm. It then discusses the challenges of using web archives such as the size of the information and archives, but also that there is not a single archive but instead a patchwork of different archiving activities collected at different times and in different ways. These archives are also subject to change over time as they are not static archives and can appear or disappear with very little notice. AI is also playing a part in more accurate automated services and helping future researches deal with huge digital archives. Use of these tools with have to go beyond the simple free text search—such as linked data. In turn, the Google type of search is not a practicable approach to dealing with large collections of emails or web archives and new research and approaches such as file linking and proper use of AI tools are needed to deal with processing this data.

Zhang, Y., Jatowt, A., Bhowmick, S. S., & Tanaka, K. (2016). The Past is Not a Foreign Country: Detecting Semantically Similar Terms across Time. IEEE Transactions on Knowledge and Data Engineering, 28(10), 2793–2807. https://doi.org/10.1109/TKDE.2016.2591008

The article works through solving the *temporal counterpart search* problem. Saying that our knowledge of the past (and its vocabulary) tends to be limited. Due to our limited knowledge of vocabulary used in the past our searching may be hampered. Ideally, users of past collections should receive some assistance when interreacting with the collections to allow the to use them as efficiently as they would be able to use current collections such as the web. This requires returning semantically similar terms from the past to an input query from the present time. This allows mapping between terms across time. They also use an extended method called *local correspondence* that locally constrains a query by transforming its core context terms, which are then automatically detected and treated as reference points. The method of the article also enhances results by outputting evidence which explain why particular terms should be considered as a temporal counterpart using principal component analysis (PCA). The article also demonstrated two effective ways for automatically finding training sets of anchor pairs for transformation matrix and proposed a method for correcting OCR driven errors as a post-processing step and introduce a new approach for explaining and visualizing results.